INTELLIGENCE INFO

ISSN 2821 - 8159, ISSN – L 2821 – 8159, Volumul 2, Numărul 2, Iunie 2023

Preocupări legislative în mineritul datelor

Nicolae Sfetcu

Sfetcu, Nicolae (2023), Preocupări legislative în mineritul datelor, *Intelligence Info*, **2**:2, 98-105, DOI: 10.58679/II26090, https://www.intelligenceinfo.org/preocupari-legislative-in-mineritul-datelor/

Publicat online: 01.05.2023

Preocupări legislative în mineritul datelor

Nicolae Sfetcu
nicolae@sfetcu.com

Legislative concerns in data mining

Abstract

Data mining involves six common classes of tasks: anomaly detection, association rule learning, clustering, classification, regression, and summarization. While the term "data mining" itself has no ethical implications, it is often associated with mining information about human behavior (ethical and otherwise). Copyright holders are directly interested in data mining issues.

Keywords: data mining, privacy, ethics, copyright, law

Rezumat

Mineritul datelor (data mining) implică șase clase comune de sarcini: detectarea anomaliilor, învățarea regulilor de asociere, clustering, clasificare, regresie, și sumarizare. În timp ce termenul de „mineritul datelor" („data mining") în sine nu are implicații etice, el este adesea asociat cu mineritul de informații în legătură cu comportamentul oamenilor (etic și de altă natură). Respectarea drepturilor de autor sunt direct interesate de problemele legate de mineritul datelor.

Cuvinte cheie: mineritul datelor, data mining, confidențialitate, etica, drepturi de autor, legislație

INTELLIGENCE INFO, Volumul 2, Numărul 2, Iunie 2023, pp. 98-105
ISSN 2821 - 8159, ISSN – L 2821 – 8159, DOI: 10.58679/II26090
URL: https://www.intelligenceinfo.org/preocupari-legislative-in-mineritul-datelor/

Introducere

Mineritul datelor implică șase clase comune de sarcini: (1)

1. Detectarea anomaliilor (detectarea valorilor anterioare / modificării / abaterii) - Identificarea înregistrărilor de date neobișnuite, care ar putea fi interesante sau erori de date care necesită investigații suplimentare.
2. Învățarea regulilor de asociere (modelarea dependenței) - Caută relații între variabile. De exemplu, un supermarket ar putea colecta date despre obiceiurile de cumpărare ale clienților. Folosind învățarea regulilor de asociere, supermarketul poate determina ce produse sunt cumpărate frecvent împreună și poate utiliza aceste informații în scopuri de marketing. Aceasta este uneori denumită analiza coșului de piață.
3. Clustering - este sarcina de a descoperi grupuri și structuri în date care sunt într-un fel sau altul „similare", fără a utiliza structuri cunoscute în date.
4. Clasificare - este sarcina de a generaliza structura cunoscută pentru a se aplica noilor date. De exemplu, un program de e-mail poate încerca să clasifice un e-mail ca „legitim" sau ca „spam".
5. Regresie - încearcă să găsească o funcție care modelează datele cu cea mai mică eroare.
6. Sumarizare - oferă o reprezentare mai compactă a setului de date, inclusiv vizualizarea și generarea de rapoarte.

Standarde

Au existat unele eforturi pentru a defini standarde pentru procesul de minerit a datelor, de exemplu Procesul standard european inter-industrial pentru extragerea datelor din 1999 (CRISP-DM 1.0) și standardul Java Data Mining din 2004 (JDM 1.0). Dezvoltarea succesorilor acestor procese (CRISP-DM 2.0 și JDM 2.0) a fost activă în 2006, dar a stagnat de atunci. JDM 2.0 a fost retras fără a ajunge la o versiune finală.

Pentru schimbul de modele de extragere - în special pentru utilizarea în analiza predictivă - standardul cheie este Predictive Model Markup Language (PMML), care este un limbaj bazat pe XML dezvoltat de Data Mining Group (DMG) și acceptat ca format de schimb de multe aplicații de minerit a datelor. După cum sugerează și numele, acesta acoperă doar modelele de predicție, o sarcină specială de minerit a datelor de mare

importanță pentru aplicațiile de afaceri. Cu toate acestea, extensii pentru a acoperi (de exemplu) gruparea sub-spațială au fost propuse independent de DMG.

Mineritul web

Conform lui (2), mineritul web și descoperirile tiparelor ascunse în cantitatea mare de date găsește informații necunoscute, relevante și utile, conținute în documentele web (3) (4). Tehnicile de minerit pe web sunt inspirate din tehnicile de minerit a datelor. Nu utilizează în mod direct tehnicile de minerit a datelor din cauza naturii diverse a datelor web care sunt disponibile sub formă de date nestructurate, semistructurate și structurate. Pentru analiza documentelor web, există mai multe sarcini de minerit și algoritmi în literatură. Spre deosebire de depozitarea de date, web are tipuri mixte de date, de ex. date de conținut (text, audio, video și grafică), date de structură (hyperlinkuri, grafuri web) și date de utilizare (date de jurnal web). Pe baza tipurilor de date utilizate, mineritul web poate fi clasificat ca mineritul conținutului web, mineritul structurii web sau analiza linkurilor și mineritul utilizării web (4).

Mineritul conținutului web descoperă informațiile utile și relevante din conținutul paginii web care ar putea fi text nestructurat, date XML, tabele structurate, informații grafice, imagini, videoclipuri etc (4). De exemplu, clasificarea documentelor web în funcție de conținutul lor, recenzii despre produse de minerit, sentimentele utilizatorilor în datele blogului.

Mineritul structurii web Se ocupă în mod special de structurile intra și inter documente, adică structura de legături a conținutului într-o pagină web și interconectivitatea paginii web între site-uri web. Structura paginii web afectează clasarea acesteia. Mineritul structurii web poate fi clasificată ca structură de hyperlink și structură de document (5). Structura de linkuri conectează conținutul în locații diferite din aceeași pagină web sau poate fi utilizată pentru a interconecta diferitele pagini web ale aceluiași site sau al unui site web diferit, în timp ce structura documentului organizează conținutul paginii sub forma structurii datorită diferitelor etichete HTML și XML.

Mineritul utilizării web descoperă modelele de traversare ale utilizatorului din jurnalele web care înregistrează fluxurile de clicuri ale utilizatorului. Mulți algoritmi de minerit a datelor sunt aplicabili și în mineritul utilizării web. Mineritul utilizării web

folosește mai mulți algoritmi de minerit a datelor. Principala problemă cu mineritul utilizărilor web sunt datele neprocesate din fluxul de clic din fișierul jurnal de utilizare a web. Mineritul web moștenește procesul utilizat în mineritul de date. Ambele diferă în ceea ce privește tehnicile de culegere de date. Datele din depozitul de date sunt colectate din diferite surse eterogene, cum ar fi fișierele plate ale bazelor de date. Acest proces implică curățarea, integrarea și transformarea datelor. Datele pentru minerit din depozitul de date sunt deja colectate, în timp ce pentru mineritul web sarcina de colectare a datelor este plictisitoare, dar cumva crawlerele web sunt utile în această activitate. După ce colectarea datelor este finalizată, aceasta necesită preprocesare, integrare, transformare și selectare a datelor necesare pentru mineritul web. În cele din urmă, se face generalizarea și analiza.

Subsarcini ale mineritului web

Mineritul web include patru sarcini secundare:

1. *Colectarea resurselor*: această fază preia documentele dorite și este realizată de motoarele de căutare web sau crawlerele web (6).
2. *Selectarea/preprocesarea informațiilor*: după găsirea resursei, documentele web relevante sunt selectate și transformate în formă standard. Majoritatea metodelor au folosit lucrări pentru a selecta datele și reprezintă datele în formă tabelară (7).
3. *Generalizare*: încearcă să afle modelul general de acces al utilizatorilor în cadrul și între site-uri web. Aceasta determină interesul și comportamentul utilizatorului. Sunt utilizate tehnicile de minerit web, cum ar fi clasificarea, tehnicile de reguli de asociere în cluster etc.
4. *Analiză/validare*: acest pas analizează, interpretează și validează informațiile potențiale în raport cu modelele de informații. Scopul acestei sarcini este mineritul cunoștințelor din informațiile obținute prin pașii anteriori. Există mai multe modele pentru a simula și valida datele web pentru minerit.

Mineritul web moștenește tehnicile de minerit de date pentru a extrage automat informațiile pentru a obține cunoștințe din conținutul web. Evaluarea modelelor implică generalizare, clasificare în cluster și analiză.

Preocupări privind confidențialitatea și etica

Conform lui (1), în timp ce termenul „mineritul datelor” („data mining”) în sine nu are implicații etice, el este adesea asociat cu mineritul de informații în legătură cu comportamentul oamenilor (etic și de altă natură).

Modalitățile în care extragerea datelor poate fi utilizată, în unele cazuri și contextele, pot ridica întrebări cu privire la confidențialitate, legalitate și etică. În special, seturile de date guvernamentale sau comerciale de extragere a datelor în scopuri de securitate națională sau de aplicare a legii, cum ar fi Programul de conștientizare totală a informațiilor sau în ADVISE, au ridicat probleme legate de confidențialitate.

Mineritul datelor necesită pregătirea datelor care pot descoperi informații sau modele care pot compromite confidențialitatea și obligațiile de confidențialitate. O modalitate obișnuită de a se produce acest lucru este prin agregarea datelor. Agregarea datelor implică combinarea datelor împreună (posibil din diverse surse) într-un mod care să faciliteze analiza (dar care ar putea, de asemenea, să facă identificarea datelor private, la nivel individual, deductibilă sau evidentă). Acesta nu este data mining în sine, ci un rezultat al pregătirii datelor înainte de - și în scopul - analizei. Amenințarea la adresa confidențialității unei persoane intră în joc atunci când datele, odată compilate, determină minerul de date sau ca oricine care are acces la setul de date nou compilat să poată identifica anumite persoane, mai ales când datele au fost inițial anonime.

Se recomandă ca o persoană să fie informată despre următoarele înainte de colectarea datelor:

- scopul colectării datelor și al oricăror proiecte (cunoscute) de minerit a datelor;
- cum vor fi utilizate datele;
- cine va putea să extragă datele și să utilizeze datele și derivatele acestora;
- starea securității în jurul accesului la date;
- cum pot fi actualizate datele colectate.

De asemenea, datele pot fi modificate pentru a *deveni* anonime, astfel încât persoanele fizice să nu fie ușor identificate. Cu toate acestea, chiar și seturile de date „de-identificate"/"anonimizate" pot conține suficiente informații pentru a permite identificarea persoanelor, așa cum s-a întâmplat atunci când jurnaliștii au reușit să găsească mai multe persoane pe baza istoriei unui set de căutare care au fost lansate din greșeală de AOL.

Dezvăluirea din neatenție a informațiilor de identificare personală care conduc la furnizor încalcă Practicile corecte de informare. Această indiscreție poate provoca vătămări financiare, emoționale sau corporale persoanei indicate. Într-un caz de încălcare a confidențialității, patronii Walgreens au intentat un proces împotriva companiei în 2011

pentru vânzarea de informații pe bază de rețetă companiilor de minerit a datelor care, la rândul lor, au furnizat datele companiilor farmaceutice.

Europa

Europa are legi destul de puternice privind confidențialitatea și se depun eforturi pentru a consolida în continuare drepturile consumatorilor. Cu toate acestea, Principiile Safe Harbor U.S.-E.U. expun în prezent efectiv utilizatorii europeni la exploatarea confidențialității de către companiile din SUA. Ca o consecință a dezvăluirii privind supravegherea globală a lui Edward Snowden, au existat mai multe discuții pentru revocarea acestui acord, deoarece, în special, datele vor fi expuse pe deplin Agenției Naționale de Securitate, iar încercările de a ajunge la un acord au eșuat.

Statele Unite

În Statele Unite, problemele legate de confidențialitate au fost abordate de Congresul SUA prin adoptarea unor controale de reglementare, cum ar fi Legea privind portabilitatea și responsabilitatea asigurărilor de sănătate (HIPAA). HIPAA cere persoanelor să-și dea „consimțământul informat" cu privire la informațiile pe care le furnizează și la utilizările prezente și viitoare ale acestora. Potrivit unui articol din *Biotech Business Week*, „[în practică], HIPAA nu poate oferi o protecție mai mare decât reglementările îndelungate din domeniul cercetării", spune AAHC. Mai important, scopul regulii de protecție prin consimțământul informat este subminat de complexitatea formelor de consimțământ care sunt solicitate pacienților și participanților, care se apropie de un nivel de incomprehensibilitate ridicat pentru indivizii medii." Acest lucru subliniază necesitatea anonimatului datelor în practicile de agregare și extragere a datelor.

Legislația S.U.A. privind confidențialitatea informațiilor, cum ar fi HIPAA și Family Educational Rights and Privacy Act (FERPA) se aplică numai domeniilor specifice pe care le abordează fiecare astfel de lege. Utilizarea mineritului de date de către majoritatea companiilor din S.U.A. nu este controlată de nicio legislație.

Legea drepturilor de autor

Europa

Din cauza lipsei de flexibilitate în legislația europeană a drepturilor de autor și a bazelor de date, mineritul datelor cu drepturi de autor, cum ar fi mineritul web, fără permisiunea proprietarului drepturilor de autor, nu este legală. Acolo unde o bază de date reprezintă date pure în Europa, probabil că nu există drepturi de autor, dar pot exista drepturi de bază de date, astfel încât exploatarea datelor devine supusă reglementărilor din Directiva privind bazele de date. La recomandarea evaluării Hargreaves, acest lucru a determinat guvernul Regatului Unit să își modifice legea drepturilor de autor în 2014 pentru a permite exploatarea conținutului ca limitare și excepție. A doua țară din lume care face acest lucru după Japonia, care a introdus o excepție în 2009 pentru mineritul de date. Cu toate acestea, din cauza restricțiilor directivei privind drepturile de autor, excepția din Regatul Unit permite extragerea de conținut numai în scopuri necomerciale. Legea drepturilor de autor din Regatul Unit nu permite, de asemenea, ca această prevedere să fie înlocuită de termenii și condițiile contractuale. Comisia Europeană a facilitat discuțiile cu părțile interesate cu privire la mineritul de text și date în 2013, sub titlul Licențe pentru Europa. Accentul pus pe soluția la această problemă legală fiind licențele și nu limitările și excepțiile, a determinat reprezentanții universităților, cercetătorilor, bibliotecilor, grupurilor societății civile și editorilor cu acces deschis să părăsească dialogul cu părțile interesate în mai 2013.

Statele Unite

Spre deosebire de Europa, natura flexibilă a legii americane privind drepturile de autor și, în special, utilizarea loială, permite extragerea de conținut în America, precum și în alte țări cu utilizare loială, cum ar fi Israel, Taiwan și Coreea de Sud, fiind considerată legală. Întrucât extragerea de conținut este transformatoare, adică nu înlocuiește opera originală, este considerată legală în condițiile utilizării loiale. De exemplu, ca parte a soluționării Google Book, judecătorul președinte al cazului a hotărât că proiectul Google de digitalizare a cărților cu drepturi de autor era legal, în parte din cauza utilizărilor transformatoare pe care le-a afișat proiectul de digitizare - una fiind extragerea de text și date.

Bibliografie

(1) Bentley, Drew (2022). *Business intelligence și analitica în afaceri*. MultiMedia Publishing, ISBN 978-606-033-779-9, Licență CC BY-SA 4.0. Traducere și adaptare: Nicolae Sfetcu, https://www.telework.ro/ro/e-books/business-intelligence-si-analitica-in-afaceri/

(2) Santosh Kumar and Ravi Kumar, ”A Study on Different Aspects of Web Mining and Research Issues”, 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1022** 012018. Licența CC BY 3.0. Traducere și adaptare: Nicolae Sfetcu

(3) Srivastava J., Cooley R., Deshpande M, Tan, P-N. Discovery and Applications of Usage Patterns from Web Data ACM SIGKDD Explorations Newsletter, 2000, 1(2) 12-23

(4) Johnson, F., Gupta, S. K., Web Content Mining Techniques: A Survey, International journal of computer applications (0975-888), vol. 47, no. 11, June 2012.

(5) Tyagi N., Gupta S.K. (2018) *Web Structure Mining Algorithms: A Survey*. In: Aggarwal V., Bhatnagar V., Mishra D. (eds) *Big Data Analytics. Advances in Intelligent Systems and Computing*, vol 654. Springer, Singapore.

(6) Crimmins, F., Smeaton, A. F., Dkaki, T. and Mothe, J. TetraFusion: information discovery on the Internet. Journal of IEEExpert, pp 55-62, July 1999.

(7) Kushmerick, N. Gleaning Answers from the Web. IEEE Intelligent Systems. Vol. 14, No. 2, pp. 20-22,1999.